Comparison between Artificial and Real Faults for Fault Localization Techniques

Group 55: Zhou Xu, Jie Song, Jian Yang, Rui Cao, Buqin Wang

Abstract:

In this article, the pearson's correlation coefficient between artificial faults and real faults is calculated , which confirms the conclusion from the reference paper (Evaluating and improving fault localization techniques, by Spencer Pearson, Jose Campos, Rene Just etc. Technical report UW-CSE-16-08-03) that real faults and artificial faults are not identical . Furthermore, cross validation on the performance of fault-localization (FL) techniques are performed. And the techniques are assessed through two metrics. Results indicate that the rank for mean EXAM score is the same as mentioned in the paper, while FLT rank shows some difference.

1. Introduction:

It is indisputable that faults would exist in almost every single program, and some of them are quite hard to identify. This situation urges developers to take advantage of specific methods to find these faults with a high efficiency. During the history of bug-fixing, some FL techniques are proposed and utilized to fix bugs.

Basically, given at least one failed test and zero or more passed test, a FLT will generate a sorted list of possible faulty statement (s) with suspicious score S(s). The score formulas for 7 different FL techniques are listed in **Fig. 1**. They can be split into two groups: one is spectrum based, the other one is mutation based.

-Spectrum Based

Tarantual $S(s) = \frac{failed(s)/totalfailed}{failed(s)/totalfailed+passed(s)/totalpassed}$

Ochiai $S(s) = \frac{failed(s)}{\sqrt{totalfailed \cdot (failed(s)+passed(s))}}$

Op2 $S(s) = failed(s) - \frac{passed(s)}{totalpassed+1}$

Barinel $S(s) = 1 - \frac{passed(s)}{passed(s)+failed(s)}$

Dstar $S(s) = \frac{failed(s)}{passed(s)+(totalfailed-failed(s))}$

-Mutation Based:

Muse $M(m) = failed(m) - \frac{f2p}{p2f} \cdot passed(m)$ $S(s) = avg_{m \in mut(s)} M(m)$

Metallaxis $M(m) = \frac{failed(m)}{\sqrt{totalfailed \cdot (failed(m)+passed(m))}}$ $S(s) = max_{m \in mut(s)} M(m)$

Fig. 1 Formula for suspiciousness of each statement or mutant

Based on Fig 1, the best FL technique can be selected among several FL techniques after scoring on faulty programs filled with faults. But these faults are usually generated by mutation tools or manually rather than real faults in a program. In the past, papers have been published on evaluating those FL techniques through artificial faults. But how comparable those artificial faults are from real faults is not very clear.

In professor Rene Just's paper, evaluation of same set of FL techniques under both artificial faults and real faults are carried out. Significant difference in ranking results indicate that real faults can't be replaced by artificial faults, which is shown in **Fig. 2.** In addition, professor Just's paper replicated the comparison between pairs of FL techniques and the results do not agree wholly with previous results. Further exploration on designing better fault localization techniques was presented with ranking results.

Artificial Faults			Real Faults		
Technique	*EXAM*	# Worse	Technique	*EXAM*	# Worse
Metallaxis	0.0197	6	DStar	0.0443	3
Op2	0.0437	5	Ochiai	0.0445	3
DStar	0.0442	4	Barinel	0.0453	2
Ochiai	0.0448	3	Tarantula	0.0477	2
Barinel	0.0503	1	Op2	0.0527	2
Tarantula	0.0512	0	Metallaxis	0.0768	1
MUSE	0.0574	0	MUSE	0.2186	0

(a) Techniques sorted by mean *EXAM* score or tournament ranking.

Artificial Faults		Real Faults	
Technique	FLT rank	Technique	FLT rank
MUSE	2.86	Metallaxis	3.35
Metallaxis	2.93	DStar	3.78
Op2	3.81	Ochiai	3.81
DStar	3.96	Op2	3.89
Ochiai	4.08	Barinel	4.07
Barinel	5.18	Tarantula	4.11
Tarantula	5.18	MUSE	4.98

(b) Techniques sorted by mean FLT rank

Fig 2. Fault Localization Techniques Sorted By Average Performance

In this paper, we are taking out empirical study on the results from professor Just's paper, trying to verify whether there is a correlation between real faults and artificial faults. Furthermore, cross validation on ranking of fault localization techniques are also carried out and compared with the results from professor Just's paper.

2.Method:

2.1 Pearson Correlation Coefficient:

The Pearson product-moment correlation coefficient is a measure of the linear dependence between two variables X and Y. It has a value between +1 and −1 inclusive, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

Pearson's correlation coefficient when applied to a sample is commonly represented by the letter r and may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient. We can obtain a formula for r by substituting estimates of the covariances and variances based on a sample into the formula above. So if we have one dataset {x1,...,xn} containing n values and another dataset {y1,...,yn} containing n values then that formula for r is:

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

where:

n, x_i, y_i are defined above,

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ the sample mean, and analogously for y_i

In our design, we take the mean score of artificial and real faults as $\{x_i\}$, $\{y_i\}$,and calculate the Pearson Correlation Coefficient in order to judge the correlation between artificial and real faults.

2.2 Repeated Random Sub-Sampling Validation

This method randomly splits the dataset into training and validation data. For each such split, a model is fitted to the training data, and predictive accuracy is assessed using the validation data. The results are then averaged over the splits. The advantage of this method is that the proportion of the training/validation split is not dependent on the number of iterations (folds). The disadvantage of this method is that some observations may never be selected in the validation subsample, whereas others may be selected more than once. In other words, validation subsets may overlap. This method also exhibits Monte Carlo variation, meaning that the results will vary if the analysis is repeated with different random splits.

As the number of random splits approaches infinity, the result of repeated random sub-sampling validation tends towards that of leave-p-out cross-validation.

In a stratified variant of this approach, the random samples are generated in such a way that the mean response value (i.e. the dependent variable in the regression) is equal in the training and testing sets. This is particularly useful if the responses are dichotomous with an unbalanced representation of the two response values in the data.

For real and artificial faults separately, the two metrics are applied, one is mean score metric, the other one is FLT rank metric (see section 2.3). Experiments details are listed in below and in **Fig. 3** as well:

1 Randomly split the scores of real or artificial faults into two sets: $\frac{2}{3}$ is training set and $\frac{1}{3}$ is testing set. And the testing data set is reusable for next testing set selection.

2 Get two sorted ranking lists of FL techniques from training and testing respectively during each validation. For each list, we examined whether the difference between two adjacent FL techniques is significant or not by doing T test and comparing the p-value. Then 1st ranked technique or techniques are picked out.

3 Compare the ranked 1st technique/techniques in those two lists. If they are same, this ranking is regarded successful. If not, another t-test would be applied to judge whether the difference is significant or not. If not significant, we still consider the ranking of training set correct; on the contrary, if the difference is significant, the ranking will be considered incorrect and discarded.

4 Repeat steps 1-3 for 100 times.

5 Choose the ranking list with highest frequency.

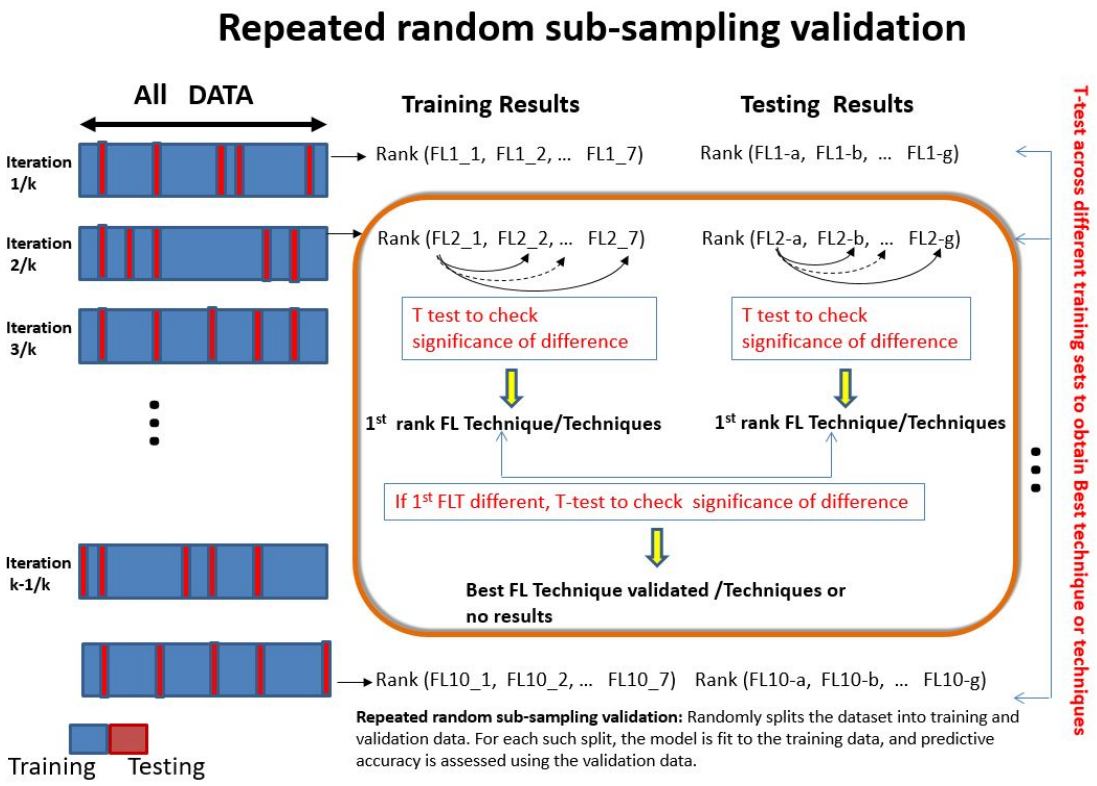

Fig.3 Repeated random sub-sampling validation

2.3 Fault Localization Techniques Evaluation Metrics

Two metrics are applied to evaluate each technique during each training-testing validation, same as described in professor Just's paper.

1: Mean EXAM Metric ---Mean EXAM score across either artificial faults or real faults

2: FLT-rank Metric ------ During each training-testing iteration, ranking for each FL technique was calculated from 1 to the total number of techniques according to their mean EXAM score. Then averaged ranking was assigned to each FLT.

3.Results:

3.1 Result of Pearson Correlation

Based on the data we have, the mean score of real and artificial faults for each FL technique is calculated and plot as in Fig. 4 .

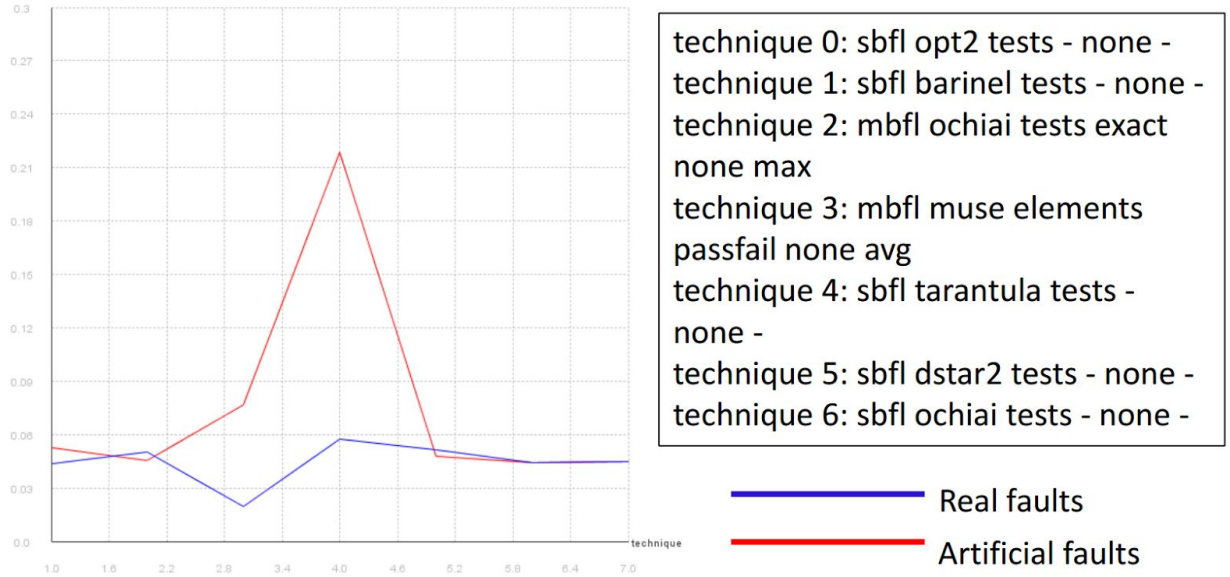

Fig. 4 Mean EXAM score of each FL technique for real and artificial faults

From the plot above, it is quite obvious that the score distributions of real and artificial faults are different, which means the correlation between real and artificial faults is very weak.

Additionally, we also calculated the Pearson's correlation coefficient. For Pearson's correlation, if it is near 1, the correlation is strong; if near 0, the correlation is weak. The result we get is

r=0.32015, much smaller than 1. Therefore, it can be concluded that real faults and artificial faults have no convincing correlation.

3.2 Cross Validation Result on real_vs_artificial Data Set.

Following tables show the ranking of FL technique performance on real faults and artificial faults respectively, with Mean Score metric and FLT Metric.. **Table 1** is for the artificial faults and **Table 2** is for real faults.

Artificial Faults					
Mean EXAM Metric			FLT Rank Metric		
Rank	Score	P-Value	Rank	Score	P-Value
Metallaxis	0.0196		Tarantula	0.023	
Op2	0.0457	1.05E-95	Dstar	0.306	2.875E-135
Dstar	0.0462	3.14E-72	Ochiai	0.359	6.420E-65
Ochiai	0.0469	2.91E-88	MUSE	1.522	5.420E-173
Barinel	0.0527	3.24E-126	Metallaxis	1.699	1.114E-68
Tarantula	0.0538	3.72E-62	Barinel	1.749	8.481E-50
MUSE	0.0591	1.22E-31	Op2	2.208	1.013E-115
Repeat time		100	Repeat time		100
Top List			Top List		
Metallaxis(98% based on p=0.05)			Tarantula(100% based on p=0.05)		

Table 1 Ranking of FL techniques on artificial faults

Real Faults					
Mean EXAM Metric			FLT Rank Metric		
Rank	Score	P-Value	Rank	Score	P-Value
Dstar	0.0498		Tarantula	0.046	
Ochiai	0.0502	1.03E-05	ochiai	0.408	1.725E-85
Barine	0.0508	1.04E-20	Dstar	0.415	1.466E-6
Tarantula	0.0523	9.03E-37	Barinel	0.954	9.898E-99
Op2	0.0575	4.79E-30	Metallaxis	1.750	1.254E-112
Metallaxis	0.0707	1.55E-48	Op2i	2.066	8.801E-68
MUSE	0.2170	6.91E-125	Muse	2.375	1.028E-53
Repeat time		100	Repeat time		100
Top List			Top List		
Dstar(33% based on p=0.05)			Tarantula(69% based on p=0.05)		

Table 2 Ranking of FL techniques on real faults

Above results shows that on artificial faults, Metallaxis performs the best according to mean EXAM while Tarantula does the best according to FLT rank

For real faults, the best performed FL technique is Dstar according to mean EXAM score, which matches with professor Just's paper; the best is Tarantula according to FLT rank, while in paper MUSE is the worst.

Our rank results from the mean EXAM metric are exactly the same as the results from the paper. Still some differences exist: for the real faults, the reference paper shows that difference between many techniques are insignificant, but our results show that the differences are significant because all p-values between neighboring techniques are larger than 0.05. This comes from the Cross Validation. We repeated 100 times for one Cross Validation, which separated each technique.

For FLT ranking metric, our results are quite different from the paper result. In our guess, the reason is that the experiments in the reference paper did not take programming error type of different bugs into consideration. For FLT rank metric, each FL technique is scored according to the rank score on each bugs. We believe that same FL technique would have different performance on different types of errors/bugs. Thus the rank of all FL techniques will be different depending on the population of mixed bug types. The reference paper's experiment mix all the bugs together and the results from the paper is the mean score of FL techniques on all the bugs. While in our experiment, we separate all bugs to training and testing set according to cross validation and the result from our experiment is mainly based on the training set of the data (testing set is mainly used to validate the results of training set), which means our results are mainly based on partial bug types. If the bug types distribution is different from training set to another training set, we will expect different results. So our suggestion is that it would be better if the source data can include the bug type information and do the experiment according to different types of bug .

For each metric, the best technique on detecting artificial and real faults varies, which means the performance of FL techniques on artificial faults cannot imply that on real faults. It agrees with our conclusion that that real faults and artificial faults have no strong correlation.

Real Faults	
MEAN SCORE METRIC	
RANK	SCORE
susp-averaging	0.1238
mrsbfl-susp-maxing	0.1239
susp-maxing	0.1239
failover	0.124
mrsbfl-failover	0.1249
mrsbfl-susp-averaging	0.1256

sbfl barinel elements	0.1267
sbfl dstar2 tests	0.1267
sbfl ochiai tests	0.1269
sbfl tarantula tests	0.1294
sbfl dstar2 elements	0.1356
sbfl muse elements	0.1406
sbfl opt2 tests	0.1419
Hybrid ochiai elements passfail mirror avg	0.1426
hybrid ochiai elements passfail mirror max	0.1427
sbfl muse tests	0.1427
hybrid ochiai tests passfail mirror max	0.1427
hybrid ochiai tests passfail mirror avg	0.143
hybrid ochiai tests all mirror max	0.143
hybrid barinel tests passfail mirror max	0.1431
hybrid barinel tests type mirror max	0.1433
hybrid ochiai elements all mirror avg	0.1433
hybrid dstar2 tests passfail mirror max	0.1434
hybrid dstar2 tests type mirror max	0.1434
hybrid ochiai tests all mirror avg	0.1434
hybrid ochiai elements all mirror max	0.1434

Table 3 Top 25 ranking of 596 FL techniques on real faults

At last, we also ranked the top 25 among the 596 techniques for real faults by applying mean score metric and 100 times repeated cross validation. **Table 3** is the result.

4. Conclusion:

By applying Pearson correlation test for real faults and artificial faults, we found the pearson correlation coefficient is 0.32015 which means real faults and artificial faults have no strong

correlation. So the performance of FL techniques on artificial faults cannot imply that on real faults. Also from our Cross Validation result on real_vs_artificial data set, we found the rank of techniques for two faults are different, which also suggests the big difference between artificial faults and real faults. We also found top 25 FL techniques based pure_real_faults data set. It suggests these top FL techniques may be useful in finding real faults. For 596 techniques real faults, our rank is different from on the paper and further investigation is still needed.

Reference: Evaluating & Improving Fault Localization Techniques, Spencer Pearson, Jose Campos, Rene Just, Gordon Fraser, Rui Abreu, Michael D. Ernst, Deric Pang, Benjamin Keller, *Technical Report UW-CSE-16-08-03,* August 2016